DOCUMENT RESUME

ED 446 149 TM 031 915

AUTHOR Kenney, Patricia Ann; Silver, Edward A.; Alacaci, Cengiz; Zawojewski, Judith S.

TITLE Content Analysis Project--State and NAEP Mathematics Assessments. Report of Results from the Maryland-NAEP Study.

INSTITUTION Pittsburgh Univ., PA. Learning Research and Development Center.

SPONS AGENCY National Assessment Governing Board, Washington, DC.
PUB DATE 1998-11-00
NOTE 61p.; For a related paper about the content analysis design features, see TM 031 916.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Comparative Analysis; *Content Analysis; *Grade 8; Junior High Schools; *Mathematics Tests; National Competency Tests; Test Construction; Test Content; Test Items; *Test Results
IDENTIFIERS Experts; *Maryland School Performance Assessment Program; *National Assessment of Educational Progress

ABSTRACT

As part of the Content Analysis Project, a panel of six experts in mathematics education examined the congruence between the Maryland grade-8 test and the National Assessment of Educational Progress (NAEP) test. Two panelists were very familiar with the Maryland School Performance Assessment Program (MSPAP), two with the NAEP, and the other two were familiar with neither test. Maryland was one of two states chosen for a similar analysis; the other, North Carolina uses an assessment composed entirely of multiple-choice questions, while the MSPAP is entirely constructed-response questions. The panel concluded that the differences found between the MSPAP and the NAEP at grade 8 along the content dimension are not sufficient to account for the magnitude of difference between proficient performance on the Maryland test and proficient performance on the NAEP. Neither are the slight differences between the Maryland test and the NAEP test at grade 8 along the cognitive dimension sufficient to account for the magnitude of differences between proficient performance on the Maryland test and proficient performance on the NAEP. The panelists offered three possible reasons for the discrepancy: (1) differences in stakes or consequences between the two tests; (2) the connections between the test and instruction; and (3) the standards-setting processes used. (Contains 12 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 446 149

CONTENT ANALYSIS PROJECT -- STATE AND NAEP MATHEMATICS ASSESSMENTS

Report of Results from the Maryland-NAEP Study

Conducted Under Contract with the
National Assessment Governing Board

Patricia Ann Kenney and Edward A. Silver, Co-Principal Investigators

Cengiz Alacaci, Post-doctoral Research Associate
Learning Research and Development Center
University of Pittsburgh

Judith S. Zawojewski, Associate Professor
National-Louis University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM031915

November 1998

BEST COPY AVAILABLE

Section 1

Design of the Study

Background

In its Redesign Policy (1996), the National Assessment Governing Board (NAGB) outlined a number of goals and objectives for guiding changes in the National Assessment of Educational Progress (NAEP). One particular set of goals and objectives involved assisting states in linking their assessments with NAEP, and several states have already begun to establish such links. In particular, some states have begun to report student performance in terms of state-level "proficiency" standards, employing language similar to that used in the NAEP achievement levels. For many states that participate in NAEP, however, discrepancies have emerged between the percentages of students scoring at the NAEP proficient level and those meeting the state standard for proficient performance (Musick, 1996; Archer, 1997). In general, the trend is that the percentage of students meeting the state standard for proficient performance as defined by the state is higher than that of students in the state NAEP sample who meet the "proficient" achievement level as defined by NAEP.

What factors contribute to these differences in proficient performance on a state's assessment and proficient performance on NAEP? There are many possible reasons for the performance differences including variations in the purposes of the assessments, in the definitions of "proficient" and the processes used to set proficiency standards, or in content coverage between the state test and NAEP. Musick (1996) proposed that it is important to examine the state assessment programs and NAEP in order to identify the possible reasons for these differences.

Based on Musick's report and on conversations with state policy makers, NAGB funded a study that would address possible reasons for the differences in performance levels and recruited Patricia Ann Kenney (Research Associate) and Edward A. Silver (Senior Scientist) from the

University of Pittsburgh's Learning Research and Development Center (LRDC) to conduct the study. Drs. Kenney and Silver were assisted by two LRDC colleagues -- Judith S. Zawojewski (Research Associate) and Cengiz Alacaci (Graduate Research Assistant). Hereafter in this report, these four people are referred to as the LRDC project staff.

The focus of the NAGB-sponsored study was on the tests[1] themselves and did not involve issues about how the proficient levels were defined and set. In particular, the study examined the congruence[2] between a state's test in eighth-grade mathematics and that used by NAEP and then used the results to answer a fundamental question: Are identified differences between the state assessment and NAEP sufficient to account for the magnitude of difference between proficient performance on the two tests? Additionally, an important part of the study was the development of a model process that states could use to compare their frameworks and assessments to NAEP not only in mathematics but also in other content areas (e.g., reading, science).

Two states, North Carolina and Maryland, participated in the study. Among the reasons for including these two particular states in the study were that representatives from each state expressed strong interest in the project, and that the assessments used in each state are quite different in a number of important ways. One important way in which the assessments differed was format: the North Carolina assessment is composed entirely of multiple-choice questions and the Maryland assessment is composed entirely of constructed-response questions. It was thought that the diversity in format as well as other ways in which the state assessments differed from each other (e.g., purpose; reporting level) would contribute to the generalizability of the content analysis process.

---

[1]In this report, the terms "test" and "assessment" are often used interchangeably, following Shepard (1994). If there is a difference between these two terms, it is one of emphasis: a test usually refers to a particular coherent test instrument; an assessment is more likely to refer to a system that involves more than one test.

[2]In mathematics, two geometric figures are said to be congruent if they can be superimposed so as to coincide, and there are a number of ways that congruence can be demonstrated mathematically. In this study, we use the word "congruence" as a synonym for the relationships between important components of each test (e.g., the congruence between the state framework and the NAEP framework). Absolute judgments about congruence were not possible, but we believed that it would be possible to describe the congruence (or lack of it) between the two tests.

Design Features

In thinking about the design of a process to investigate the congruence between a state's test and NAEP, the LRDC project staff considered three features to be especially important: 1) the process should involve consensus judgments by a panel whose members were selected on the basis of their expertise in areas relevant to the study (e.g., middle school mathematics, the state test, NAEP); 2) the state test and NAEP should be examined from multiple perspectives (e.g., technical characteristics; content areas; cognitive demand) and according to an array of aspects (e.g., test frameworks and specifications; test items; scoring guides and student work); and 3) the process should be multi-phased (i.e., there should be adequate time for the LRDC project staff to prepare materials and analyze data and for the panelists to discuss important issues and to reach consensus). Each of these design features is discussed next.

The Consensus Judgments of a Panel of Experts

In recent years, basing decisions about NAEP on expert judgment has become a common occurrence. For example, expert judgment about student performance is at the heart of the achievement levels-setting process (NAGB, 1990). For the NAEP mathematics assessment, judgments of mathematics education professionals were used to establish the content and curricular validity of the tests that comprised the trial state assessments in 1990 and 1992 (Silver & Kenney, 1994; Silver, Kenney, & Salmon-Cox, 1992) and to examine the 1992 NAEP achievement levels-setting efforts (Silver & Kenney, 1993). For this study, we used a panel of experts to assist in making the congruence judgments for each state assessment and NAEP.

The panel of experts charged with examining the relationship between a state's assessment and NAEP was composed of six mathematics education professionals (e.g., mathematics teachers, college/university mathematics educators, mathematics curriculum specialists), and the composition of the panel reflected distributed expertise that spanned the state test, NAEP, and middle school mathematics. Of the six members, two members were selected on the basis of their familiarity with the state assessment; that is, they served in a capacity that ensured knowledge of the state's testing program (e.g., serving on the mathematics framework development committee; writing test items;

providing professional development for mathematics teachers on the state assessment program). Personnel from the state's department of education nominated possible panelists, and the LRDC project staff contacted them. Having representatives from the state as members of the panel ensured that states were an integral part of the content analysis process. Also, the two "state" panelists served as resource people when informational questions arose about the state test.

Another pair of panelists were selected on the basis of their knowledge of the NAEP mathematics assessment, and in particular the NAEP grade-8 test. For example, these panelists had served on committees that developed the NAEP mathematics framework and items, or they knew about NAEP through their involvement with other NAEP-related projects such as the NCTM NAEP Interpretive Reports Project (Kenney & Silver, 1997; Silver & Kenney, in press). The two "NAEP" panelists could provide the group with expertise about that test, should the need arise.

The last two panelists were selected for their expertise about and experience with middle school mathematics education and for their lack of specialized knowledge about either the state assessment or about NAEP. The role of these panelists within the group was one of neutrality with respect to the tests to be examined; that is, this pair of "neutral" panelists had no vested interest in either test.

The six panelists met for two days in two separate sets of meetings held about a month apart. The structure of the panel -- two state panelists, two NAEP panelists, two neutral panelists -- allowed varying points of view to emerge during the discussions. Also, in instances where the panelists would work in small groups, it was possible to form two subgroups of three members, (one NAEP, one state, one neutral). It is important to note here that different six-member panels were selected for each state-NAEP comparison; that is, no panelists served on both the North Carolina and Maryland panels. This was done to ensure that direct comparisons would not be made between the state tests, but only about the state test and NAEP.

In addition to the six members of the panel, there were others who participated in the consensus process and who brought with them additional expertise to the consensus process. For example, state testing directors, testing consultants, and mathematics specialists from the states

were invited to participate in the activities and discussions at the meetings. Members of the NAGB staff and a member of NAGB (Mark Musick) also were involved in some of the deliberations. Representatives from the National Center for Education Statistics (NCES) also attended some meetings. Finally, the LRDC project staff members, in addition to providing additional expertise about NAEP, were responsible for creating all materials used at the meetings, analyzing data, and serving as facilitators of the panelists' and other participants' discussions about the congruence between the tests.

The Tests as Viewed from Multiple Perspectives

In formulating the design for this study, we proposed that tests could be compared according to multiple perspectives, hereafter referred to as "dimensions." Three dimensions common to the state test and NAEP were identified as relevant to this study. First, there is a technical dimension that involves components such as the number and type of items, the time allotted to administering the test, the difficulty of the items, etc. A second aspect involves a content dimension that has to do with the particular content topics (e.g., for mathematics -- geometry, measurement, algebra) included. And a cognitive dimension involves the extent to which a test engages students in various cognitive processes, including problem solving, reasoning, or the recall of facts and definitions. Because each test has a distinct profile with respect to these dimensions, it is possible to determine the profile for each test and then to compare the tests for congruence on all three dimensions. A model for this process appears in Figure 1.1. The methods used to examine the technical, content, and cognitive aspects of each test are described briefly next. Specific examples of the methods and materials used to examine the dimensions can be found in the state reports for North Carolina (Section 2) and Maryland (Section 3) and in the Procedural Appendix, which is a separate document from this report.

Technical dimension. The technical aspect involves particulars such as the number of items on the test, the type of items (multiple choice; constructed response), the time allotted for students to take the test, the difficulty of the items, etc. Information about a test's technical characteristics most often appears in documents such as frameworks and specifications and the testing program's

technical reports. It was deemed important to this study that the panelists and other participants be knowledgeable about the technical aspects of the state assessment and NAEP before beginning their comparison of the tests themselves.

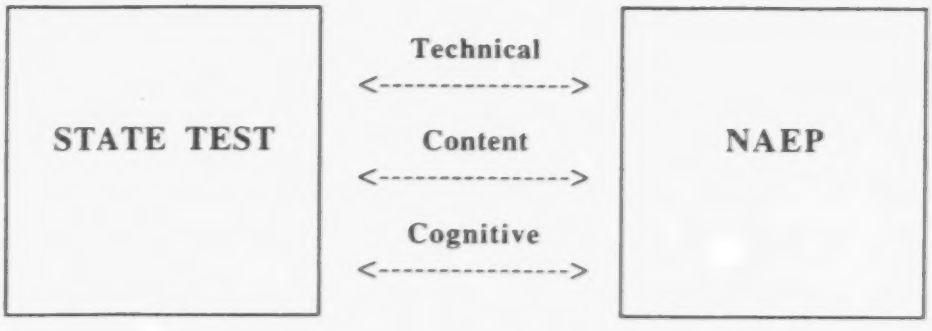


Figure 1.1. Dimensions along which to investigate congruence between the state test and NAEP

In this study, information about the technical aspects of the state assessment and that of NAEP was obtained in two ways. First, the LRDC project staff, using the framework documents and technical reports, prepared summaries of technical information about each test to be used at the panel meetings and in the project's final report. These summaries were verified for accuracy by the representative from the state's department of education and by members of the NAGB staff.

The second way that technical information was obtained was through presentations made at the first meeting. In particular, a representative from the state's department of education presented an overview of the purpose of the state test and its important technical characteristics; a member of the LRDC project staff who was very familiar with NAEP gave a similar presentation about that assessment, and the NAGB staff members provided additional information when necessary. These presentations enabled the panelists and other participants to get further clarification on the technical aspects of each test.

Content dimension. The content aspect of a test has to do with what is being assessed; that is, in mathematics this involves the particular topics included on the test (e.g., number concepts and relationships; measurement; geometry; statistics and probability; algebra). These content topics are common to most mathematics assessments, but the content coverage across topics can vary widely depending on the purpose of the test. For example, on a basic competency test a large percentage of the items could be devoted to topics in number properties and computation, with lower percentages of items in the other content areas. Another grade-8 test, which has a more broadly defined purpose, could include items that are equally distributed across content areas. Thus, for the tests just described, although both include the same content topics, the coverage is different, thus likely affecting the content congruence between the tests.

The content aspects of a state test and NAEP and the congruence between them were investigated in two ways: a framework-to-framework matching by content area and an item-to-framework cross-matching of items from one test onto the framework of the other test. The framework-to-framework matching activity involved comparing carefully the content topics as presented in the 1996 NAEP mathematics framework document (The College Board, 1994) and those presented in the relevant mathematics "framework" document from the state. For a state, the relevant framework document is most likely the state's curricular goals for mathematics at each grade level or clusters of grade levels (e.g., grades 6-8), and there is evidence that many states use the curricular goals as the test specifications for their testing programs (Roeber, Bond, & Braskamp, 1997). Because both the NAEP mathematics framework and the state curricular goals are based in large part on content topics, using these documents in the framework-to-framework activity was deemed reasonable. The activity itself involved the panelists and other participants identifying topics in both the NAEP framework and the state framework that were similar, topics that were in NAEP but missing in the state framework, and topics that were in the state framework but missing in NAEP. Comparing the common topics in each framework and the topics unique to each framework provided a way to evaluate the congruence between NAEP and the state test on the basis of intended content coverage, as specified in the frameworks.

The item-to-framework, cross-matching activity involved having the panelists and other participants classify NAEP items according to the content topics in a state's framework document and items from the state test according to the NAEP framework. This activity was designed to serve two purposes. First, it provided an additional opportunity to examine the items from each test. Additionally, the results from the activity can be used to validate the information from the framework-to-framework matching through "triangulation" of the data. In qualitative research, triangulation is a standard technique that draws on multiple methods and data sources to gain more confidence in the accuracy of the findings (Jick, 1983; Mathison, 1988). For example, in the context of this study, suppose that the framework-to-framework matching activity revealed that the measurement topic of converting units within the same system (e.g., inches to feet; millimeters to centimeters) appears in both the NAEP mathematics framework and in the state's framework document. Then, it should be highly likely that in the item-to-framework cross-matching activity, NAEP items classified as assessing conversion of units should be classified in the state framework's category associated with conversion of units, and vice versa. If the expected classification occurred, then the framework-to-framework match was confirmed. However, if the item classification went in another direction (e.g., the item was classified in an unexpected category), then the outcome would be "non-confirmation" of the framework-to-framework match, and reasons for this non-confirmation could be explored.

Results from the two activities just described allowed the panelists and other participants to evaluate the congruence between the NAEP and state tests with respect to their content dimension. Additional information about the congruence came from discussions among the panelists and participants as they completed each activity. The LRDC project staff members served as the facilitators of these discussions.

Cognitive dimension. The cognitive aspect of a test refers to the extent to which a test engages students in various cognitive processes such as recalling important facts and definitions, computing with numbers, demonstrating conceptual understanding, and using reasoning in mathematical situations. In designing this study, we recognized the importance of comparing the

tests with respect to the cognitive demands each test placed on students, the premise being that even though two tests might be similar in terms of what content topics are included, they could be quite different on how the topics were assessed. How topics are assessed on the test goes well beyond content area and item format considerations and into the realm of whether the focus is on lower-order skills such as recall of facts and routine procedures or on higher-order skills such as problem solving and mathematical reasoning, or a combination of both kinds of skills.

The cognitive aspects of the NAEP and the state test were compared on the basis of two activities. First, the LRDC project staff chose a set of criteria external to both assessments that could be used to evaluate the cognitive demand of the items on each test. The criteria were obtained from sources such as the Curriculum and Evaluation Standards for School Mathematics (NCTM, 1989) and other studies involving NAEP (e.g., Romberg, Smith, Smith, & Wilson, 1992; Silver & Kenney, 1994), and the criteria represented both high-level (problem solving, communication, reasoning) and low-level (recall of facts, routine procedures) cognitive processes. The panelists and other participants used these criteria to evaluate items from NAEP and the state assessment. The findings from this part of the investigation were based on the results of the evaluation of the items according to cognitive demand and on a discussion among the panelists facilitated by the LRDC project staff members.

Because the Maryland assessment program uses a test composed completely of constructed-response items, it was important for the comparison between that test and NAEP to focus on the cognitive demand of those items along with their scoring guides and sample student responses at each score level. In particular, the design of this activity was based on this idea: If the cognitive demand of the item is high, then is that high cognitive demand sustained in the scoring guide for that item and in the set of sample student responses for each score level? The panelists and other participants had the opportunity to examine carefully some constructed-response items from NAEP and from the Maryland test. The issues concerning the cognitive demand of the items and whether that demand was sustained in the scoring guides and student work were then discussed by the group.

The Multi-phase Process

Because the process involved expert judgment concerning the congruence of the state test and NAEP along multiple dimensions, it was important to plan carefully the sequence of events so that the panelists would have adequate time to examine each test completely, to discuss important issues as a group, and to reach consensus. Also, the LRDC project staff needed time to analyze the data generated by the panelists, to synthesize the results of the group discussions, and to plan ways in which to share information with the panelists so as to inform their judgments. Based on these considerations, it was decided that the process should be multi-phased, with five distinct phases: pre-meetings, first panel meeting, between meetings, second panel meeting, and post-meetings.

The chart in Figure 1.2 outlines the five phases and contains a brief summary of the activities occurring in each phase. Two of the five phases (Phases II and IV) involved the activities occurring during the two-day meetings of the panel of experts. In general, considering the technical and content aspects of the state test and NAEP was the focus of the first meeting; examining the cognitive aspects of the tests was the focus of the second meeting, with the final portion of that meeting devoted to a discussion of the question concerning whether differences in the technical, content, or cognitive dimensions were sufficient to account for the performance differences at the proficient level between the state test and NAEP. The other three phases (Phases I, III, and V) allowed LRDC project staff to obtain and study relevant documents and other materials from NAEP and from the state assessment, to prepare materials for the meetings, to analyze data generated during the meetings, and to produce summaries for the panel meetings and the final report.

Phase I: Before the First Meeting

- LRDC project staff gathered information on the state test and NAEP and prepared handouts and focus questions to be sent to the panelists.
- LRDC project staff compiled information on the technical aspects of each test.
- LRDC project staff prepared activities for the first meeting concerning the tests' content aspects.
- Panelists read handouts and responded to the focus questions.

Phase II: First Meeting

Day 1

- Representative from the state and from NAEP gave presentations on their respective assessment programs.
- Panelists discussed their responses to the focus questions completed prior to the meeting; LRDC project staff served as facilitators for this discussion.
- Panelists worked individually and then in small groups on the framework-to-framework matching activity (as described on page 1-7); LRDC project staff served as facilitators in the small group discussions.

Day 2

- Panelists discussed the findings from the framework-to-framework matching activity and reached consensus on the congruence between the two tests based on content characteristics.
- Panelists worked individually on the item-to-framework cross-matching activity (NAEP items to state framework; state items to NAEP framework - as described on pages 1-7 and 1-8)

Phase III: Between the Meetings

- LRDC project staff analyzed the results of the item-to-framework matching activity to validate the judgments of the panelists from the framework-to-framework matching activity.
- LRDC project staff prepared a summary of the content congruence decisions to share with the panelists.
- LRDC project staff compiled information about criteria that was used to evaluate the cognitive aspects of the tests and sent it to the panelists; project staff also prepared materials for use during the second meeting.
- Panelists received and read information about the criteria to be used at the meeting to evaluate the cognitive aspects of the tests.

Phase IV: Second Meeting

Day 1

- LRDC project staff shared findings from the activities completed at the last meeting and facilitated a discussion by the panelists on the content congruence between the state test and NAEP.
- Panelists worked individually on an activity that asked them to evaluate the cognitive demands of a set of NAEP items and a set of items from the state test.
- LRDC project staff produced a preliminary analysis of the data from the cognitive demand activity for presentation at Day 2 of the meeting.

Figure 1.2. The phases of the study and brief summary of the activities within each phase.

Phase IV: Second Meeting (continued)

Day 2

- LRDC project staff shared the preliminary findings from the cognitive demand activity with the panelists and facilitated a discussion of those findings.
- [for the Maryland-NAEP meetings]. Panelists completed an activity concerning the level of cognitive demand as sustained from constructed-response item to scoring guide to examples of student work at each score level. LRDC project staff facilitated the discussion based on this activity.
- Panelists engaged in a discussion, facilitated by the project staff, concerning the congruence between the tests on their cognitive characteristics.
- Based on their judgments about the congruence between the tests on the three dimensions (technical, content, cognitive), the panelists worked to reach consensus on the differences between the state test and NAEP and whether these differences were sufficient to account for the magnitude of difference between proficient performance. Panelists also suggested other factors that could be contributing to the performance differences.

Phase V: After the Meetings

- LRDC project staff prepared a report that summarized the findings from the project and submitted that report to the state and to NAGB.

Figure 1.2. The phases of the study and brief summary of the activities within each phase. (continued)

Limitations of the Study

As stated previously, the purpose of this study was to examine the congruence between a state's test in eighth-grade mathematics and that use by NAEP to answer a fundamental question: Are identified differences between the state assessment and NAEP sufficient to account for the magnitude of difference between proficient performance on the two tests? This specificity of purpose imposed these limitations on the study:

1) There was no direct attempt to evaluate either the state assessment or NAEP as a part of this study. Instead, we assumed that the state assessment was carefully developed and had undergone some kind of evaluation, and we looked for documentation (e.g., technical reports, research studies) that supported these assumptions. In the case of NAEP, there is evidence that it has been extensively evaluated by external groups such as the National Academy of Education (1992, 1994) and more recently the National Academy of Sciences (1998).

2) The study's design stopped short of comparing the tests according to the ways in which results were reported. Our charge was to consider only the tests themselves according to technical, content, and cognitive characteristics and for constructed-response questions, the way in which such questions were scored. Comparing the tests according to reporting issues might be the focus of another study.

3) The study was concerned exclusively with the subject area of mathematics. Although the LRDC project staff made a conscious attempt to use design principles that could be applied to subject areas other than mathematics (e.g., reading; science), the extent to which the design principles actually can be used to evaluate the congruence between tests in other subject areas should be established in future studies.

Concluding Comments

In this section, we have summarized background information concerning the perceived need for this study and presented an explanation of and rationale for the key features of the study design. The design features of consensus judgment by a panel of experts, examination of the state test and NAEP along multiple dimensions, and the organization of the process into multiple phases were selected as not only being relevant to the study of the congruence between two mathematics tests, but also because it was thought that these features would generalize to other content areas and grade levels assessed by NAEP and by states. For example, with regard to the panel of experts, the members can be selected according to their expertise in other disciplines such as reading or science and according to their expertise at the elementary, middle school, or high school levels. And it is likely that any state test and the NAEP test in a discipline other than mathematics can be evaluated along the three dimensions -- technical, content, and cognitive -- described in this section, although the details would vary by discipline. The multi-phase process, which is discipline-independent, serves as a suggested structure for the study itself. These three design features, then, can contribute to the development of a model process for examining the congruence

between a state test and NAEP that can be used in disciplines other than mathematics and at grade levels other than grade 8. We also presented some important limitations to the study.

In the next section of this report, we present details of how the design described here was implemented in two studies involving the state mathematics assessments at grade 8 in two states: North Carolina and Maryland. A separate document contains additional information about the model process and suggestions on how to implement that process.

References

Archer, J. (1997, January 15). States struggle to ensure data make the grade. Education Week, pp. 1, 30.

The College Board. (1994). Mathematics framework for the 1996 National Assessment of Educational Progress. Washington, D.C.: National Assessment Governing Board.

Jick, T. J. (1983). Mixing qualitative and quantitative methods: Triangulation in action. In J. V. Maanen (Ed.), Qualitative methodology. Beverly Hills, CA: Sage.

Kenney, P. A., & Silver, E. A. (Eds.) (1997). Results from the Sixth Mathematics Assessment of the National Assessment of Educational Progress. Reston, VA: National Council of Teachers of Mathematics.

Mathison, S. (1988). Why triangulate? Educational Researcher, 17(2), 13-17.

Musick, M. D. (1996, June). Setting educational standards high enough. Atlanta: Southern Regional Education Board.

National Academy of Education. (1992). Assessing student achievement in the states. Stanford, CA: The Academy.

National Academy of Education. (1994). The Trial State Assessment: Prospects and realities. Stanford, CA: The Academy.

National Academy of Sciences. (1998). Grading the Nation's Report Card: Evaluating NAEP and transforming the assessment of educational progress. Washington, D. C.: National Research Council.

National Assessment Governing Board. (1996, August). Policy statement on redesigning the National Assessment of Educational Progress. Washington, D.C.: Author.

National Assessment Governing Board. (1990, May). Setting appropriate achievement levels for the National Assessment of Educational Progress. Washington, D. C.: Author.

National Council of Teachers of Mathematics. (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: Author.

Roeber, E., Bond, L., & Braskamp, D. (1997). Annual survey of state student assessment programs, Fall 1996. Washington, D.C.: Council of Chief State School Officers.

Romberg, T. A., Smith, M., Smith, S., & Wilson, L. (1992, June). The feasibility of using international data to set achievement levels for the National Assessment of Educational Progress (NAEP). Madison, WI: National Center for Research in Mathematics Education.

Shepard, L. (1994). The challenge of assessing young children. Phi Delta Kappan, 76(3), 206-210.

Silver, E. A., & Kenney, P. A. (1993). Expert panel review of the 1992 NAEP Mathematics Achievement Levels. In National Academy of Education, Setting performance standards for student achievement: Background studies (pp. 215-282). Stanford, CA: The Academy.

Silver, E. A., & Kenney, P. A. (1994). The content and curricular validity of the 1992 NAEP Trial State Assessment (TSA) in mathematics. In National Academy of Education, The Trial State Assessment: Prospects and realities: Background studies (pp. 231-284). Stanford, CA: The Academy.

Silver, E. A., & Kenney, P. A. (Eds.) (In press). Results from the Seventh Mathematics Assessment of the National Assessment of Educational Progress. Reston, VA: National Council of Teachers of Mathematics.

Silver, E. A., Kenney, P. A., & Salmon-Cox, L. (1992). The content and curricular validity of the 1990 NAEP Mathematics Items: A retrospective analysis. In National Academy of Education, Assessing student achievement in the states: Background studies (pp. 157-218). Stanford, CA: The Academy.

SECTION 2

THE MARYLAND - NAEP STUDY

SUMMARY

As part of the Content Analysis Project, funded by the National Assessment Governing Board, a panel of six experts in mathematics education examined the congruence between the Maryland grade-8 test and the NAEP grade-8 test. The panel members were selected so that their expertise was distributed according to familiarity with the two tests. Specifically, two panelists were very familiar with the Maryland School Performance Assessment Program (MSPAP), two panelists with NAEP, and the remaining two panelists were familiar with neither test. In addition to the panelists, additional expertise was provided by the LRDC project staff and representatives from NAGB and from the National Center for Education Statistics (NCES).

The tests used in this study were both administered in 1996. They were examined from multiple perspectives and along the technical, content, and cognitive dimensions. The consensus judgments of the panelists about the congruence between the tests became the basis for answering a fundamental question: Are identified differences between the Maryland test in eighth-grade mathematics and that used by NAEP sufficient to account for the magnitude of difference between proficient performance on the two tests -- 48% on the Maryland test at grade 8 (1994-96 results) and 24% on state-NAEP at grade 8 (1996 results) (Musick, 1996)?

The process of evaluating the congruence between the Maryland test and NAEP involved a five-phase process and a variety of activities. Two of the five phases (Phases II and IV) involved the activities occurring during the two-day meetings of the panel of experts and other participants. In general, considering the technical and content aspects of the state test and NAEP was the focus of the first meeting (Phase II), held on February 4-5, 1998; examining the cognitive aspects of the

tests was the focus of the second meeting (Phase IV), held on March 26-27, 1998, with the final portion of that meeting devoted to a discussion of the question concerning whether differences in the technical, content, or cognitive dimensions were sufficient to account for the performance differences at the proficient level between the state test and NAEP. The other three phases (Phases I, III, and V) allowed LRDC project staff to obtain and study relevant documents and other materials from NAEP and from the state assessment, prepare materials for the meetings, analyze data generated during the meetings, and produce summaries for the panel meetings and the final report.

Based on the results from activities completed by the panelists and other participants and on their discussions, two specific conclusions, one conclusion concerning the content congruence and one conclusion about cognitive congruence, and one general conclusion were agreed upon by the panelists:

Conclusion: Congruence along the Content Dimension

There are differences between the Maryland test and the NAEP test at grade 8 along the content dimension. However, these differences are not sufficient to account for the magnitude of difference between proficient performance on the Maryland test and proficient performance on NAEP.

Conclusion: Congruence along the Cognitive Dimension

There are only slight differences between the Maryland test and the NAEP test at grade 8 along the cognitive dimension. These slight differences are not sufficient to account for the magnitude of difference between proficient performance on the Maryland test and proficient performance on NAEP.

Overall Conclusion

There are differences between the Maryland test at grade 8 and the NAEP test at that grade level, but the differences are not sufficient to account for the magnitude of the difference between proficient performance on the Maryland test and proficient performance on NAEP.

Introduction

The next subsections of the Content Analysis Project report contain a summary of the Maryland - NAEP study and key findings concerning the congruence between the Maryland School Performance Assessment Program (MSPAP) in grade-8 mathematics (hereafter referred to as the Maryland grade-8 test or MSPAP) and the NAEP grade-8 mathematics assessment (hereafter referred to as the NAEP grade-8 test or NAEP). Below is a brief description of the contents of each subsection.

Subsection

2.1 *Participants in the Maryland-NAEP Meetings*

Contains information about the panel of experts who participated in the study and about the other participants (e.g., representatives from the Maryland State Department of Education and from NAGB).

2.2 *Schedule of Meetings and Related Activities*

Contains a description and chronology of the activities completed prior to, between, and during the meetings of the panel of experts.

2.3 *Learning about the Maryland School Performance Assessment Program (MSPAP) and NAEP*

Contains information about the activities designed to inform the panelists and other participants about important aspects of MSPAP and NAEP.

2.4 *Comparing the Maryland and NAEP Tests Along the Technical Dimension*

Contains a summary of how the two tests compared on selected technical aspects such as the number of items, item format, and level of difficulty.

2.5 *Comparing the Maryland and NAEP Tests Along the Content Dimension*

Contains a summary of the activities that the panelists and other participants completed and their decisions concerning the congruence between the tests along the content dimension.

2.6 *Comparing the Maryland and NAEP Tests Along the Cognitive Dimension*

Contains a summary of the activities that the panelists and other participants completed and their decisions concerning the congruence between the tests along the cognitive dimension.

2.7 *Concluding Comments*

Contains a summary of comments made by the panelists and other participants concerning the tests themselves and about the procedures used in the study to compare them.

2.8 *Reference List*

A list of published sources that the LRDC project staff used to inform their work on the Maryland-NAEP study.

2.1: Participants in the Maryland-NAEP Meetings

Examining the congruence between the Maryland (MSPAP) grade-8 test and the NAEP grade-8 test involved the consensus judgment of a panel of experts who were specifically chosen to participate in the deliberations. Figure 2.1 contains information about each panelist, and for the state and NAEP panelists a brief description of their involvement with the particular assessment. The composition of the panel was as follows: two panelists (Anita G. Morris, Arthur Rodriguez) familiar with MSPAP, two panelists (Glendon W. Blume, Diana Wearne) with expertise in NAEP, and two panelists (Donald Chambers, Frances R. Curcio) who were not familiar with either assessment but with expertise in mathematics education. The LRDC project staff, in consultation with the NAGB staff, selected the two panelists familiar with NAEP and the two "neutral" panelists, but the nominations for the panelists familiar with the Maryland test came directly from representatives from the Maryland State Department of Education.

In addition to the panelists, these other people observed or participated in the Maryland-NAEP deliberations: William Schafer (Director, Student Assessment Division) and Cindy Hannon (Mathematics Specialist) from the Maryland State Department of Education; members of the NAGB staff (in particular, Mary Crovo and Mary Lyn Bourque); and representatives from NCES (in particular, Suzanne Tripplett). The LRDC project staff members (Patricia Ann Kenney, Edward A. Silver, Judith S. Zawojewski, and Cengiz Alacaci) were responsible for creating all materials used at the meetings, documenting the discussions at the meetings, and facilitating the consensus process based on the expert judgment of the panelists and the other attendees who chose to participate. Together, the set of participants brought with it a high degree of distributed expertise about MSPAP, NAEP, and middle school mathematics education, thus contributing to a rich discussion of the issues involved in comparing the two assessments.

Panelists Familiar with MSPAP	
Anita G. Morris Coordinator of Mathematics Anne Arunel County Public Schools Annapolis, MD 15+ years of experience as a mathematics teacher, content specialist; coordinates county-wide activities related to MSPAP	Arthur Rodriguez Mathematics and Computer Science Teacher Great Mills High School (MD) 20+ years of experience as a mathematics/computer science teacher; developed performance tasks in mathematics from MSPAP for every year of testing
Panelists Familiar with NAEP	
Glendon W. Blume Associate Professor Department of Curriculum and Instruction The Pennsylvania State University Member of the 1996 NAEP mathematics framework committee; author of book chapters and journal articles based on NAEP results	Diana Wearne Associate Professor School of Education University of Delaware Mathematics educator and researcher; author of book chapters based on NAEP mathematics results
"Neutral" Panelists	
Donald Chambers Educational Consultant / State Mathematics Supervisor for Wisconsin (retired) Madison Wisconsin Currently serving as a consultant with the Wisconsin Center for Educational Research	Frances R. Curcio Associate Professor Mathematics Education New York University Mathematics educator and researcher; author of numerous publications; active in promoting mathematics education internationally

Figure 2.1. Members of the Expert Panel for the Maryland-NAEP study.

2.2: Schedule of Meetings and Related Activities

The multi-phased process of examining the congruence between the Maryland grade-8 test and the NAEP grade-8 test according to three dimensions (technical, content, cognitive) involved two meetings of the panelists and the other participants and three time periods devoted to related activities. Figure 2.2 contains a description and chronology of the activities completed prior to, between, and during the meetings. More detailed descriptions of the activities and discussions are included in the sections of this report that follow. In general, however, the first meeting and the time prior to it was devoted to issues concerning the technical and some content aspects of the two assessments; the time between meetings and the first part of the second meeting focused on the remaining content aspects; and the latter part of the second meeting was devoted to examining the cognitive characteristics of the tests, reaching consensus on the congruence between the assessment and deciding whether the differences identified were sufficient to account for the magnitude of the difference in proficient performance.

Phase I: Before the First Meeting

- LRDC project staff gathered information on the Maryland test and NAEP and prepared handouts and focus questions to be sent to the panelists.
- LRDC project staff compiled information on the technical aspects of each test for distribution at the first meeting.
- LRDC project staff prepared activities for the first meeting based on the tests' content aspects.
- Panelists read handouts and responded to the focus questions.

Phase II: First Meeting February 2-3, 1998

Day 1

- Representatives from Maryland and from NAEP gave presentations on their respective assessment programs.
- Panelists discussed their responses to the focus questions completed prior to the meeting; LRDC project staff served as facilitators for this discussion.
- Panelists worked individually and then in small groups on a framework-to-framework matching activity designed to examine the content congruence between the tests; LRDC project staff served as facilitators in the small group discussions.

Figure 2.2. Summary of activities related to the Maryland-NAEP study

Phase II: First Meeting: February 2-3, 1998 (continued)

Day 2

- Panelists discussed the findings from the framework-to-framework matching activity and reached consensus on the congruence between the two tests based on content characteristics.
- Panelists worked individually on another activity concerned with the content aspects of the tests: matching the items from one test to the framework of the other test (item-to-framework cross-matching); the first part of the activity concerned with matching MSPAP items to the NAEP framework, was completed during this time.

Phase III: Between the Meetings

- Panelists completed the item-to-framework cross-matching by matching the NAEP items to the MSPAP learning outcomes and suboutcomes. Results were returned to LRDC.
- LRDC project staff analyzed the results of the item-to-framework matching activity to validate the judgments of the panelists from the framework-to-framework matching activity.
- LRDC project staff prepared a summary of the content congruence decisions to share with the panelists.
- LRDC project staff compiled information about criteria that can be used to evaluate the cognitive aspects of the tests and sent it to the panelists; project staff also prepared materials for use during the second meeting.
- Panelists received and read information about the criteria to be used at the meeting to evaluate the cognitive aspects of the tests.

Phase IV: Second Meeting: March 26-27, 1998

Day 1

- LRDC project staff shares findings from the activities completed at the last meeting and facilitates a final discussion by the panelists on the content congruence between the Maryland test and NAEP.
- Panelists work individually on an activity that asked them to evaluate the cognitive demand of a set of NAEP items and a set of MSPAP items by matching the item to categories representing various levels of cognitive demand (e.g., recall of facts, reasoning).
- LRDC project staff produces a preliminary analysis of the data from the cognitive demand activity for presentation at Day 2 of the meeting.

Day 2

- LRDC project staff shared the preliminary findings from the cognitive demand activity with the panelists and facilitated a discussion of those findings.
- Panelists further examined the cognitive demands of MSPAP and NAEP by examining constructed-response items, scoring guides and student work, and discussed their reactions.
- Panelists engaged in a discussion, facilitated by the LRDC project staff, concerning the congruence between the tests on their cognitive characteristics.
- Based on their judgments about the congruence between the tests on the three dimensions (technical, content, cognitive), the panelists worked to reach consensus on the differences between the MSPAP and NAEP and whether these differences were sufficient to account for the magnitude of difference between proficient performance. Panelists also suggested other factors that may be contributing to the performance differences.

Phase V: After the Meetings

- LRDC project staff prepared a report that summarizes the findings from the Maryland-NAEP study and submitted that report to the state and to NAGB.

Figure 2.2. Summary of activities related to the Maryland-NAEP study (continued)

2.3: Learning About the Maryland School Performance Assessment Program (MSPAP) and NAEP

Before the panelists and other participants began the process of examining the Maryland grade-8 test and the NAEP grade-8 test along technical, content, and cognitive dimensions, it was important that they become familiar with important aspects of each assessment program. The project design provided a number of opportunities for them to learn about and understand important aspects of each test including the test frameworks and specifications and the kinds of items included on the test.

One way for the panelists to become familiar with the Maryland test and NAEP involved reading source documents such as test framework and specifications, technical reports, and the state's curriculum guidelines, which in many cases also serve as the framework and specifications for the test. Prior to the first meeting, the panelists were sent the 1996 NAEP framework document (The College Board, 1994) and the Maryland learning outcomes and suboutcomes for mathematics (Maryland State Department of Education, 1997), the curriculum framework that also serves as the MSPAP assessment framework. To guide the panelists in their reading of these documents, the LRDC staff prepared a series of focus questions that panelists were asked to answer in writing. Examples of the Maryland document and focus questions are in Appendix A-1, and those for the NAEP assessment are in Appendix A-2. These appendices can be found in Section 3 of this report.

Additional information about each assessment was subsequently presented at the first panel meeting. In particular, William Schafer and Cindy Hannon gave a presentation about MSPAP with a focus on the grade-8 mathematics, and Patricia Ann Kenney gave a similar presentation about NAEP.

In addition to information about the Maryland and NAEP test frameworks, the panelists had an opportunity to become acquainted with the kinds of items (i.e., questions) on each test and

the relationship between the test items and the content and process categories as recommended in the framework documents. Both the Maryland and NAEP frameworks specified the content areas to be covered, such as number properties and operations, measurement, and geometry. In addition, both testing programs specified that the items assess a variety of cognitive processes and not just low-level skills. In NAEP, the processes are called "mathematical abilities" and include Conceptual Understanding, Procedural Knowledge and Problem Solving; for MSPAP the process outcomes consist of the four cross-cutting themes from the NCTM Curriculum and Evaluation Standards for School Mathematics (1989): Problem Solving, Communication, Reasoning, and Connections.

To gain an understanding of the relationship between the test items and the framework categories, the panelists were asked to complete an activity which involved matching a set of 22 released NAEP items from the grade-8 test to the NAEP framework content and ability categories. Similarly, they also matched the set of six items within a released MSPAP grade-8 task[1] (called Birth Dates) to the MSPAP mathematics outcomes and suboutcomes. An example of this activity for each test appears in Appendix B-1 (MSPAP) and Appendix B-2 (NAEP). These appendices can be found in Section 3 of this report. The panelists were provided with the official classifications for the items in the Birth Dates task and for the NAEP items so that they could check their classifications against those assigned by the MSPAP and NAEP test developers. Because the purpose of this activity was to acquaint the panelists with the frameworks and items for MSPAP and NAEP, the classifications were not further analyzed as data for the study.

The panelists and other participants had time during the first meeting to ask questions and to share their comments about each assessment. As a result of this discussion, the panelists identified and commented on some potentially important differences between the tests. However, at this point in the deliberations, panelists and other participants were not asked to make evaluative

---

[1]MSPAP differs from most other assessments in that there is a planned interrelatedness between the items; that is, the items within a task are organized around a theme. For the Birth Dates task, the four items (some of which have multiple parts) all involve the topic of birthday months and the ways in which the month in which you were born might affect preferences such as favorite color.

judgments about the possible impact that these differences could have on the performance at the proficient level between the Maryland test and NAEP. These differences are summarized below:

- The purposes of the Maryland and NAEP tests are quite different. In particular, MSPAP is designed as a criterion-referenced test that measures what students have learned based on a set of desired outcomes (The Maryland School Performance Assessment Program Learning Outcomes: Mathematics [1997]). NAEP is designed as a general survey of what U. S. students know and can do in mathematics.

- The Maryland test is aligned with the curriculum as specified in the MSPAP learning outcomes for mathematics. NAEP is not tied directly to any curriculum framework, but instead is a broad-based assessment of topics in the mathematics curriculum at grades 4, 8, and 12.

- Both the Maryland and NAEP test use matrix sampling, but the approaches are different. In MSPAP, all students in grade 8 are assessed, but no student takes the entire test. Instead, students are randomly assigned to a testing group, and they take one of three test forms (called "clusters). NAEP uses a representative probability sample based on students within schools within geographic areas. Thus, all students in grade 8 take one form of MSPAP; not all eighth-grade students in the U.S. take the NAEP grade-8 test, and no one student in the NAEP sample takes the entire NAEP grade-8 test.

- The primary focus of the results from the Maryland test is the school, but individual student scores are available upon request. MSPAP results are based on scale scores and on proficiency levels that describe what students can do in relation to the Maryland learning outcomes. NAEP results are reported for the nation and demographic subgroups and for states that voluntarily participate in the state-level assessment. The results are also reported in terms of scale scores, but the NAEP achievement levels are reported in terms of what students should be able to do.

- NAEP is generally perceived as a low-stakes test, and no obvious consequences are attached to performance. However, there are consequences for schools associated with performance on the Maryland test. In particular, the Maryland State Department of Education produces score reports for each school system and school consisting of

information about the percentage of students at each proficiency level. These reports become part of the state's program that rewards schools that are making substantial progress toward achieving state standards for school performance (Maryland State Department of Education, 1997) and are also used in the decision-making process regarding whether a school is eligible for reconstitution due to the lack of progress toward meeting the standards (Maryland State Department of Education, 1998).

During the discussion of the activity that had panelists classify test items according to the content areas and process categories, the panelists commented that they initially had difficulty matching the item to a particular category or categories especially in the case of the ability categories for NAEP and the process categories for MSPAP. However, the panelists agreed among themselves that the matching became easier near the end of the activity. While they did not always agree with the official classifications, they understood why a particular classification was assigned to an item.

The initial reaction of the panelists to the thematic structure of the Birth Dates task was that they preferred it over the discreteness of the items within the NAEP blocks. Also, allowing the MSPAP items to have multiple process outcomes and content outcomes and suboutcomes was viewed positively. However, some panelists raised a concern about whether linking the items thematically could possibly restrict the range of mathematical concepts being assessed, thus restricting the content coverage of the MSPAP test at grade 8. For example, the Birth Dates task appeared to support questions in the content areas of number and data analysis, statistics, and probability, but it might not support the development of items in content area such as geometry and measurement. Representatives from the Maryland State Department of Education commented that during task development, special attention is paid to choosing a variety of contexts that, when taken together, support items that span the content areas identified in the MSPAP learning outcomes.

In sum, completing a set of activities prior to the first meeting, listening to presentations about the Maryland test and NAEP at the first meeting, and engaging in a discussion of important

aspects of the tests afforded the panelists and other participants the opportunity to obtain a shared understanding of tests. This shared understanding served as a foundation for the examination of the tests along technical, content, and cognitive dimensions.

2.4: Comparing the Maryland and NAEP Tests Along the Technical Dimension

The first phase of the process of examining and comparing the Maryland grade-8 test and the NAEP grade-8 test focused on the technical aspects of each assessment. In particular, the technical characteristics of a test involved components such as the number of items, item format, time allotted for administration, the content coverage of the test, and special features such the interrelatedness of items and whether they were administered at more than one grade level. The analysis of the Maryland test and NAEP along the technical dimension was compiled by the LRDC project staff members based on information contained in official documents such as the MSPAP technical report (Maryland State Department of Public Instruction, CTB/McGraw Hill, & Measurement Incorporated, 1996) and the 1996 NAEP mathematics framework document (The College Board, 1994). The results of this analysis were shared with the panelists and other participants as needed during the meetings.

Figure 2.4 contains a summary of how the Maryland and NAEP tests at grade-8 compared on selected technical aspects, and notable differences appear in every category. For example, the MSPAP test consists entirely of constructed-response tasks, with the number of tasks varying between test forms. In NAEP, which consists of both multiple-choice and constructed-response items, each student takes three blocks of items (i.e., intact item sets) with a 15-minute time limit per block for a total of 45-minutes of testing time. Also, some NAEP items are administered at multiple grade levels to gauge the change in student performance across grades, but no MSPAP items are administered across grades.

With respect to the content coverage, both the Maryland test and NAEP include nearly the same content areas, with some differences in the structure of the content areas. The MSPAP content outcomes separate statistics and probability into two categories; in NAEP they are combined with data analysis into a single content strand -- Data Analysis, Statistics, and

Probability. The opposite situation occurs for the areas of measurement and geometry: in NAEP, these are separate categories; in MSPAP, they are included in a single outcome -- Measurement/Geometry. While NAEP specifies a particular distribution for items according to the five content strands (e.g., 15% Measurement; 25% Algebra and Functions), no specific distribution is specified in MSPAP for the mathematics content outcomes. Instead, the mathematics test developers insure that all outcomes are covered across the three test clusters developed for an assessment year.

One of the most striking differences between MSPAP and NAEP is the way in which items are related to one another. In NAEP, the vast majority of the items are discrete; that is, there is no obvious relationship between each item. Occasionally, two or three contiguous items in a NAEP block are related in that they share a common table or graph or that the answer to one item depends on the answer to a previous item. These NAEP items form item "families." In MSPAP, items are related in at least two ways. All items within a task share a common theme such as students' birth dates or the powerful forces of nature, and in many cases within a task the items depend on answers to previous items.

These differences and the others as shown in the figure suggest that along the technical dimension, the Maryland test and NAEP appear to be quite different. During their deliberations, the panelists and other participants became cognizant of the discrepancies in the technical characteristics of the two tests, but they were not asked to make a congruence judgment based on the technical dimension because by design these congruence judgments were withheld until the panelists had the opportunity to view the tests along the content and cognitive dimensions.

	MSPAP GRADE-8 TEST	NAEP GRADE-8 TEST
Item type	Constructed-response, with some items assessing knowledge in multiple subjects (e.g., science and mathematics; mathematics and writing)	Multiple choice (approx. 55% of total number of items); Constructed response (approx. 45% of total number of items)
Distribution of items by content area	• Number Concepts and Relationships (computation/ estimation, number systems and number theory; number and number relationships) • Measurement/Geometry • Statistics • Probability • Patterns/Algebra (patterns and function; algebra) All content outcomes covered, but no specific percentages for content outcomes/suboutcomes	25% Number Sense, Properties and Operations 15% Measurement 20% Geometry/Spatial Sense 15% Data, Statistics, Probability 25% Algebra and Functions
Number of items each student takes	Varies across the test clusters (i.e., non-parallel forms); usually 1 to 3 mathematics tasks per cluster, with each task consisting of multiple parts ("items")	Varies; each student takes 3 blocks of items, but number of items may not be the same for each block
Time	1 hour, 45 minutes testing time per day; 5 days for the entire MSPAP grade-level assessment; at least 30 minutes allotted for each mathematics task	45 minutes total (15 minutes for each of 3 blocks)
Relationship between items	All items within a task organized around a common theme (e.g., powerful forces of nature; birth dates); answers to some items depend on answers to previous items	Limited; a few item pairs share a common chart or graphic or in which the answer to one item depends on the answer to a previous item
Items administered at multiple grade levels	No	Yes: subset administered at grades 4 and 8; grades 8 and 12; and all three grades

Figure 2.4. Comparison of the tests according to selected technical aspects.

2.5: Comparing the Maryland and NAEP Tests Along the Content Dimension

The second phase of the process of examining and comparing the Maryland grade-8 test and the NAEP grade-8 test focused on the content characteristics of the test. In particular, the content characteristics involve what is assessed on the test; that is, the mathematical topics and the coverage of each topic on the test itself. The panelists, along with some other participants who chose to complete the activities, examined content aspects in two ways. First, they completed an activity (hereafter referred to as the "framework-to-framework matching" activity) which required matching the NAEP framework topics and subtopics for grade 8 within the five content strands to the Maryland content outcomes and suboutcomes for mathematics at grade 8 . For the second activity (hereafter referred to as the "item-to-framework cross-matching" activity), the panelists and other participants were asked to classify a set of NAEP items according to the Maryland learning outcomes and a set of items from six MSPAP mathematics tasks according to the NAEP content topics and subtopics. Each of these activities and the panelists' findings associated with each activity is discussed next.

Framework-to-Framework Matching Activity

Design. The framework-to-framework matching activity was designed around the mathematical content areas used in the Maryland grade-8 test and the NAEP grade-8 test. For this activity, the five MSPAP mathematics outcomes (Number Concepts and Relationships; Measurement/Geometry; Statistics; Probability; Patterns and Algebra) and the corresponding suboutcomes were considered as the content framework categories; for NAEP the five content strands (Number Sense, Properties, and Operations; Measurement; Geometry and Spatial Sense; Data Analysis, Statistics, and Probability; Algebra and Functions) as defined in the framework document were used.

Because of the close match between the MSPAP content outcomes and NAEP content strands, there was no need for any special restructuring of the frameworks. It was decided that panelists would be instructed to consider all of the MSPAP content outcomes and suboutcomes as they searched for topics that matched a particular NAEP content strand-topic. Instructing the panelists to consider all MSPAP outcomes and suboutcomes could ensure identification of topics that might be in both frameworks but in different content areas (for example, ordered pairs as a number topic in MSPAP as opposed to ordered pairs as an algebra topic in NAEP).

An issue had to be resolved concerning the structure of the NAEP content strands into numbered topics and lettered subtopics. For example, in the NAEP Geometry content strand for grade 8, topic #6 "Apply geometric properties and relationships in solving problems" is subdivided into three subtopics: *a.* Use concepts of 'between,' 'inside,' 'on,' and 'outside'; *b.* Use the Pythagorean relationship to solve problems; and *c.* Apply properties of ratio and proportion with respect to similarity. A question arose about whether the panelists should work at the subtopic level as they matched the NAEP framework to the MSPAP learning outcomes and suboutcomes or should they work at the topic level, using the subtopics to clarify the main topic. It was decided to have the panelists in the Maryland-NAEP study try to work at the topic level, using the subtopics to clarify the main topics.

Materials. The LRDC project staff created materials for the framework-to-framework matching activity. A sample of these materials for the content area of geometry appears in Appendix C, which can be found in Section 3 of this report. The materials for the other four content areas were similar to the set shown in the appendix.

Procedures. The six panelists were divided in two subgroups of three, with each subgroup having one person familiar with MSPAP, one person familiar with NAEP, and one neutral panelist. The other participants were invited to join a small group and complete the activity, if they chose to do so, and the LRDC project staff served as facilitators.

In each small group, the panelists were instructed first to work individually to evaluate each content area and then to meet and discuss the match between the frameworks with respect to that

content area. Their decisions were recorded on flip-chart pages. Once both small groups had discussed all five content areas, then both groups reconvened to reach a consensus judgment on the relationship between each content area and on the similarities and differences between the content aspects of the tests based on the framework descriptions.

The structure just described was chosen because it allowed for consensus-building throughout the process and for cross-validation of the decisions made by an independent group. First, within the small groups the panelists had the opportunity to discuss their individual decisions and to come to consensus about the relationship between the content areas in the MSPAP and NAEP frameworks. Then, when the two groups reconvened, not only was there another opportunity to build whole-group consensus about the relationship between the framework categories, but there also was the opportunity to cross-validate the decisions made by the two groups; that is, to examine the similarities and differences between the independent decisions made about each of the five content areas before making the compromises and changes needed to reach consensus. This model of cross-validation using two groups working independently on the same task is similar to that used to produce the scale-anchor descriptions for the 1990 NAEP mathematics assessments (Mullis, Dossey, Owen, & Phillips, 1991).

Analysis of data from the framework-to-framework matching activity. The data generated from this activity involved the decisions of the panelists as recorded on the flip-chart pages. The decisions of each small group were compared and discussed by the panelists and other participants during a large group meeting. The LRDC project staff members facilitated the discussion and took notes.

Once the two small groups reconvened to discuss the match between the Maryland and NAEP frameworks for grade 8, it did not take them long to reach consensus. In fact, the independent agreement between the two groups was remarkably high. Any major disagreements were adjudicated during the discussion of each content area, and it was relatively easy for the groups to resolve the disagreements. The LRDC project staff summarized the panelists' findings,

based on meeting notes and materials from the meeting, and then presented the findings to the panelists at a subsequent meeting for their comments and reactions.

Findings. The panelists and other participants who completed this activity reached consensus based on professional judgment about the congruence between the frameworks. The consensus was reached during a group discussion of the findings from the framework-to-framework matching activity. **The consensus was that in general there was moderate congruence with respect to the content characteristics of the Maryland and NAEP grade-8 tests based on the frameworks.** Due in large part to time constraints, there was no systematic effort to have the panelists judge the congruence between the individual content areas in MSPAP and NAEP (e.g., between Algebra and Functions in NAEP and Patterns/Algebra in MSPAP). Instead, the panelists and other participants were asked to comment on why they decided that the overall congruence between frameworks was moderate.

The panelists and other participants agreed among themselves that across all five content areas, the mathematics topics as described in the MSPAP mathematics outcomes and suboutcomes and the NAEP content framework were similar, but the similarity between MSPAP and NAEP was more evident in some content areas than in others. For example, topics within the content area of measurement were nearly identical between NAEP and MSPAP. However, although the NAEP content strand Data Analysis, Statistics, and Probability and MSPAP outcomes for Statistics (#7) and Probability (#8) included similar topics, the NAEP content strand included topics that were not in the MSPAP outcomes such as sampling, randomness, and bias in data collection and counting techniques to determine the number of ways an event can occur. Overall, of the five common content areas, there were 9 NAEP subtopics (out of a total of 45 subtopics across the five content strands) identified as being in the NAEP framework but not in the MSPAP outcomes and suboutcomes; there were only 3 MSPAP suboutcomes (out of a total of 47 suboutcomes across the five mathematics outcomes) that were judged not to be in the NAEP framework. Thus, there appear to be some differences in content coverage between NAEP and MSPAP, with more NAEP topics missing from MSPAP than MSPAP topics missing from NAEP.

Another factor contributing to the judgment that the congruence was moderate had to do with the "grain size" of descriptions of the NAEP content topics and subtopics and the MSPAP mathematics outcomes and suboutcomes. Each framework contains both general and specific descriptions, but in general the panelists and other participants agreed among themselves that many of the descriptions in MSPAP are more specific than those in NAEP. For example, a MSPAP suboutcome in the Statistics outcome (#7) speaks specifically to "construct[ing] circle graphs" and another suboutcome in the Geometry outcome (#6) mentions "angle relationships formed by transversals and parallel lines."

There are good reasons for the two discrepancies just noted -- coverage differences and issue of specificity vs. generality -- that have to do with the purpose of and particulars associated with each assessment. With respect to coverage, some topics assessed in MSPAP at an earlier grade level are not repeated in the outcomes and suboutcomes for subsequent grades. For example, in MSPAP number theory concepts (e.g., even and odd numbers; factors and multiples) are assessed at grade 5 and are not repeated in the outcomes and suboutcomes for grade 8. In NAEP, however, topics often overlap between grade levels in order to provide information about performance across grade levels. Using again the number theory example, number theory topics are assessed in NAEP at all three grade levels. Thus, given these differences in the inclusion of topics at a particular grade level, it is not surprising that there are differences in content coverage between MSPAP and NAEP at grade 8.

Differences in generality and specificity of content topics can also be explained by examining the purposes of MSPAP and NAEP and how those purposes affect the respective frameworks. The MSPAP mathematics outcomes and suboutcomes represent important concepts to be included in the eighth-grade mathematics curriculum. Schools are held accountable for their students' understanding of the outcomes and suboutcomes. Thus, because the MSPAP outcomes and suboutcomes serve to guide both assessment and instruction at grade 8, it stands to reason that they must be written in specific, discrete terms. The NAEP framework also serves as a guide to the development of an assessment, but it has no direct link to any particular curriculum. Instead,

the NAEP content strands represent a broad-based survey of important mathematical topics in grade 8 and also topics that overlap between grade levels.

Item-to-Framework Cross-Matching Activity

Design. This activity was designed to serve two purposes. First, it provided yet another way to examine the content aspects of the Maryland and NAEP tests at grade 8, this time using the frameworks and the items developed according to the framework descriptions. Additionally, results from the activity could be used to validate the information from the framework-to-framework matching through "triangulation," a technique that draws on multiple data sources to gain more confidence in the accuracy of findings from qualitative research.

Although the main focus of this activity was still on examining the tests along the content dimension, the panelists and other participants were asked to classify the items from one assessment according to the process categories from the other assessment's framework. It was thought that having the panelists also classify the items with respect to categories that represented cognitive demands (i.e., Conceptual Understanding, Procedural Knowledge, and Problem Solving in NAEP; Problem Solving, Communication, Reasoning, Connections in MSPAP) could provide useful information about the cognitive dimension of the tests.

Six tasks from the 1996 MSPAP at grade 8 were used in this activity, and these six tasks were determined to represent a sample of the suboutcomes from the five MSPAP content learning outcomes.[2] The six tasks were distributed into packets so that each packet contained four tasks: two tasks common to all packets and two of the remaining four tasks. The panelists and other attendees who chose to participate worked individually to classify about 60 items according to the NAEP content and ability categories.

Similarly, nine blocks of NAEP items were selected so that the items within the blocks represented the percentage distribution across content strand topics as recommended in the NAEP framework document (e.g., about 25% of the items were classified by NAEP as Number Sense, Properties, and Operations, about 15% as Measurement, etc.). The nine blocks were distributed

[2]In each assessment year, the intent of MSPAP is to sample the suboutcomes across all mathematics learning outcomes rather than to include all suboutcomes.

into packets so that each packet contained six blocks: three blocks common to all packets and three of the remaining six blocks. Within this structure, then, each panelist classified about 70 items according to the MSPAP process outcomes and the mathematics content outcomes and suboutcomes.

Materials. The LRDC project staff created materials for the item-to-framework cross-matching activity. Sample worksheets are in Appendix D-1 for Maryland and Appendix D-2 for NAEP. These appendices are in Section 3 of this report.

Procedures. The plan for the item-to-framework cross-matching activity was to have the panelists and other attendees who chose to participate worked on this activity individually during the latter part of the first meeting. Each person would first classify items in each of the four MSPAP tasks according to the NAEP content and ability categories and then classify the NAEP items according to the MSPAP mathematics outcomes and suboutcomes. The culmination of the activity would be a discussion of the congruence between the items from one test and the framework of the other test, with the LRDC project staff serving as facilitators for the discussion and recorders of the panelists' comments.

Unfortunately, due to the amount of time devoted to the framework-to-framework matching activity and the subsequent discussion, the panelists could only complete the item-to-framework cross-matching activity for the MSPAP items and the NAEP framework. It was agreed that the panelists would complete the matching activity for the NAEP items to the MSPAP mathematics outcomes and suboutcomes during the time between the meetings, with the discussion occurring during the first part of the second meeting.

Analysis of data from the activity. The data from the item-to-framework matching activity were analyzed by the LRDC project staff during the time between meetings. The primary method of analysis involved linking the cross-classifications to the framework-to-framework matching activity so as to determine whether the cross-classification data "confirmed" the judgments made about the relationship between framework content categories. For example, suppose that in the judgment of the panelists, the content topic of similarity appears in both the frameworks. Then, it

should be highly likely that, if given NAEP items classified as assessing similarity, a majority of the panelists would classify those items according to the Maryland mathematics outcomes and suboutcomes. If this situation did not occur (e.g., the item was classified in another, unexpected category or a majority of the panelists could not agree on the most appropriate content category), then the outcome would be "non-confirmation" of the framework match, and reasons for this non-confirmation could be explored.

A limitation of this method was that it could be applied only to those content topics within a framework for which items were available. An in-house examination of the NAEP and Maryland grade-8 tests administered in 1996 revealed there were items for 34 of the 45 NAEP content topics (76%) and for 27 of the 47 Maryland mathematics outcomes and suboutcomes (57%). The fact that there were some content topics in both NAEP and MSPAP for which there were not items with official classifications in those topics had a reasonable explanation: it is the policy in both NAEP and MSPAP to sample the content topics and subtopics or mathematics outcomes and suboutcomes rather than to create a test that measures all topics or outcomes included in the framework document.

Despite the lack of items for some content topics in each assessment, the fact that there were items for nearly three-fourths of the NAEP content topics and over half of the Maryland learning outcomes suggested that it would be possible to examine confirmation for some content topics in each assessment. The decision was made to restrict the analysis to the subset of 34 NAEP content topics and subset of 27 MSPAP content outcomes and suboutcomes for which items were available.

Findings based on analysis of data from the item-to-framework cross-matching. There were three outcomes from the use of the method to link the data from the item-to-framework cross-matching activity and the framework-to-framework matching activity: 1) confirmation of the match between content framework topics; 2) confirmation of the absence of a particular topic; and 3) explainable non-confirmation; that is, the attempt to understand and explain unexpected results.

With respect to confirmation, of the 34 content topics for grade 8 within the NAEP content strands, confirmation of their match with particular MSPAP mathematics outcomes and suboutcomes was obtained for 23 topics (68%). Of the 27 MSPAP outcomes and suboutcomes for grade 8, confirmation of their match to particular NAEP content topics was obtained for 19 outcomes (70%). This finding suggests that for content topics for which there were items on each assessment, there were many topics common to both frameworks.

An example of a released NAEP item for which the panelists' classification suppo[illegible] confirmation of the results from the framework-to-framework matching activity appears in Figure 2.5.1. The reasons for deciding that the classification confirmed the framework-to-framework result are given in the figure.

Results from framework-to-framework matching activity

A moderate match between NAEP algebra and functions topic #1 (AF-1: Describe, extend, interpolate, transform, and create a wide variety of patterns and functional relationships) and the MSPAP suboutcome for the patterns part of #9 Patterns/Algebra (Generalize a relationship from a pattern, graph, or table; and given a relationship, represent it by a pattern graph, or table).

Released NAEP item classified as AF-1

From any vertex of a 4-sided polygon, 1 diagonal can be drawn.
From any vertex of a 5-sided polygon, 2 diagonals can be drawn.
From any vertex of a 6-sided polygon, 3 diagonals can be drawn.
From any vertex of a 7-sided polygon, 4 diagonals can be drawn.

How many diagonals can be drawn from any vertex of a 20-sided polygon?

Answer: _____

Results from matching the above NAEP item to the Maryland framework

75% percent agreement that the item matched the MSPAP suboutcome concerning patterns; average degree of match was 3.0 (on a scale 5 - strong to 1 - weak).

Figure 2.5.1. Example of a released NAEP item for which the classification confirmed a result from the framework-to-framework matching activity.

With respect to confirmation of absence, 5 of the 34 NAEP content topics (15%) were confirmed as missing from the Maryland framework at grade 8, but for Maryland, none of the 27 MSPAP outcomes and suboutcomes for which there were items were confirmed as missing from the NAEP framework. Of the five NAEP topics that were confirmed as absent from the MSPAP framework, there was evidence that four topics -- number theory concepts; conversion of units in measurement; intersection of geometric figures; sample spaces in probability -- were assessed in MSPAP at earlier grades, and therefore were purposely not repeated in the outcomes and suboutcomes for grade 8. The fourth topic confirmed as missing from the MSPAP outcomes and suboutcomes involved sampling, randomness and bias in data collection. An example of a NAEP item for which the panelists' classifications supported the confirmation of absence decision appears in Figure 2.5.2.

Results from framework-to-framework matching activity

Agreement that NAEP Data Analysis, Statistics, and Probability topic #3 (DA-3: Understand and apply sampling, randomness, and bias in data collection) is missing from the MSPAP mathematics outcomes and suboutcomes

Released NAEP item classified as DA-3

A poll is being taken at Baker Junior High School to determine whether to change the school mascot. Which of the following would be the best place to find a sample of students to interview that would be most representative of the entire student body?

A. An algebra class
B. The cafeteria
C. The guidance office
D. A French class
E. The faculty room

Results from matching the NAEP item to the MSPAP outcomes and suboutcomes

88% agreement that this NAEP item did not match any MSPAP outcome or suboutcome.

Figure 2.5.2. Example of a released NAEP item for which the classification confirmed the absence of the NAEP content topics in the MSPAP mathematics outcomes and suboutcomes.

With respect to explainable non-confirmation, the remaining 6 NAEP content topics (18%) and the remaining 7 Maryland competency goals (30%) showed evidence of non-confirmation. When the reasons for the non-confirmation were discussed by either the LRDC project staff, by the participants at the second meeting, or both, all 13 instances of non-confirmation could be reasonably explained. In particular, two major categories of explainable non-confirmation were identified:

- **Generality of the NAEP framework descriptions and the specificity of the MSPAP outcomes and suboutcome descriptions**

  When one description was general and the other specific, as it was in many cases between the NAEP and MSPAP frameworks, the panelists also found it difficult to match them, and if a match was identified, that match was evaluated as moderate or weak. Then, as before, when an item classified according to a particular MSPAP mathematics outcome or suboutcome had to be matched to the NAEP content strand (or vise versa), there was no consensus among the majority of the panelists about an appropriate match.

- **Conceptual density of particular items from each test**

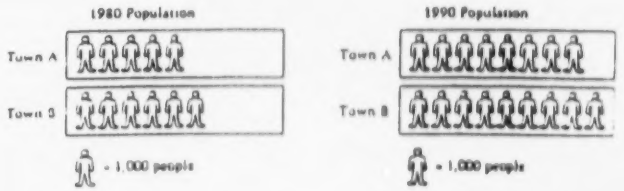
  Within a particular test item, conceptual density occurs when mathematical concepts are embedded side-by-side with other mathematical ideas in ways that are so integrated that one particular concept or idea may not take precedent over all others in the item (BOSUN Project, 1997; Zawojewski & Silver, 1998). Thus, the item can be considered as conceptually dense, and as such, it might be difficult to reach a majority agreement that any <u>one</u> concept is central to the item.

An example of a NAEP item identified by the consensus judgment of the panelists as being conceptually dense appears in Figure 2.5.3. When asked to match this item, which had an official NAEP classification in the subtopic of ratio and proportion in the content strand Number Sense, Properties, and Operations, to the primary MSPAP outcomes and suboutcomes, the panelists could not agree on the best classification. Unfortunately, because the MSPAP items used in this study are still secure, no examples of conceptually-dense MSPAP items can be shown.

Results from framework-to-framework matching activity

A strong match between NAEP Number Sense, Properties, and Operations topic #5 (NPO-5: Apply ratios and proportional thinking in a variety of situations) and the MSPAP mathematics outcomes and suboutcomes concerned with ratios and proportions in #5 (Number Concepts and Relationships: Given a problem, write the appropriate proportion and solve it; Apply ratios and proportions).

Released NAEP item classified as NPO-5



In 1980, the populations of Town A and Town B were 5,000 and 6,000, respectively. The 1990 populations of Town A and Town B were 8,000 and 9,000 respectively.

Brian claims that from 1980 to 1990 the populations of the two towns grew by the same amount. Use mathematics to explain how Brian might have justified his claim.

Darlene claims that from 1980 to 1990 the population of Town A had grown more. Use mathematics to explain how Darlene might have justified her claim.

Results from matching the NAEP item to the MSPAP outcomes and suboutcomes

Half of the panelists and other participants agreed that the item matched the MSPAP Number and Relationships suboutcome "Apply ratios and proportions." The other half agreed that the item matched the MSPAP suboutcome in Statistics "Use data analysis to write an evaluative argument in a real life situation." In each case, the degree of match between the NAEP item and the MSPAP suboutcome was about 4.0 (on a scale of 5 - strong to 1 - weak).

Decision: Explainable non-confirmation

The NAEP item is conceptually dense because it involved important mathematical concepts from number relationships and from data analysis. In particular, it requires students to read and understand data presented in a pictograph, use ratios to interpret the data, and then to write an explanation. Based on professional judgment, the panelists and other participants could not agree on one particular mathematical topic that was central to the item. Instead, they independently chose two concepts that made sense, given the conceptual density of the item.

Figure 2.5.3. An example of explainable non-confirmation due to the conceptual density of the NAEP item.

Conclusions about the Content Aspects of the Tests

Upon completion of the two of activities concerning the content characteristics of the Maryland test and the NAEP test at grade 8 (the framework-to-framework matching; the item-to-framework cross-matching) and discussing the findings from each, the panelists and other participants completed a brief questionnaire about the overall measure of the degree of congruence of the tests along the content dimension and how confident they were of this judgment. The questionnaire appears in Appendix E, found in Section 3 of this report. At this point in their deliberations, the panelists agreed that there was a moderate degree of congruence between the tests (average rating was 3.25 on a 5-point scale), and the level of confidence about that judgment was slightly above the moderate level (average rating was 3.6 on a 5-point scale). With respect to other information needed to inform their judgments about the congruence between the tests, the panelists noted that it would be important to examine the tests carefully according to their cognitive demands. These conclusions were noted:

- **There are differences between the Maryland test and the NAEP tests at grade 8 along the content dimension.**

  In particular, there are differences in generality and specificity of descriptions of content topics; that is, in general the MSPAP outcomes and suboutcomes are more specific and the NAEP framework more general. The tests also differ in the inclusion of particular topics with more topics included in NAEP than in the Maryland test. The panelists agreed that these differences are not surprising, given the differing purposes of each test.

- **However, these differences are not sufficient to account for the magnitude of difference between proficient performance on the MSPAP test and proficient performance on NAEP.**

2.6: Comparing the Maryland and NAEP Tests Along the Cognitive Dimension

The third and final phase of the process of examining and comparing the Maryland grade-8 test and the NAEP grade-8 test focused on their similarities and differences along the cognitive dimension. In this project, the cognitive dimension of the test refers to the degree to which it measures a student's ability to engage in various mathematical processes including execution of procedures, recall of facts, conceptual understanding, and problem solving. It is important to examine the tests according to the cognitive demand criteria because even though the two tests could be similar in terms of what content topics appear on the tests, they could be quite different on how that content is assessed. How topics are assessed on the test goes beyond considerations of item format (e.g., multiple choice vs. constructed response). For example, is the cognitive focus of the test mainly on recall of facts and routine procedures or on problem solving and mathematical reasoning, or a combination of both?

The panelists and other participants examined cognitive dimension of each assessment in three ways. First, during their analysis of the content dimension of the two tests, the panelists and other participants also classified a set of MSPAP items according to the NAEP ability categories (Conceptual Understanding, Procedural Knowledge, Problem Solving) and a set of NAEP items according to the MSPAP process outcomes (Problem Solving, Communication, Reasoning, Connections). During the second meeting, the panelists evaluated a subset of items from each test according to external criteria representing a variety of cognitive demand levels. Finally, because the MSPAP consists entirely of constructed-response items and because 45% of the NAEP items are also constructed-response items , it was deemed important to spend some time looking at examples of those items, their scoring guides, and student responses representing each score level.

Item-to-Framework Cross-Matching Activity: Ability and Process Categories

As reported in Section 2.3 of this report, the panelists and other participants, while classifying items from one test according to the content framework of the other test, also classified items according to categories that represented the cognitive dimension. The set of NAEP items were classified by the MSPAP process outcomes of Problem Solving, Communication, Reasoning and Connections; the set of MSPAP items were classified by the NAEP ability categories of conceptual understanding, procedural knowledge, and problem solving. If the item did not match any process or ability category, then the panelists and other participants had the option of choosing "none."

In order to obtain preliminary information about each test along the cognitive dimension, the LRDC project staff compiled the results from this activity based a simple majority of panelists' judgments about the appropriate process classification for each item. The average degree of match between the item and the process category was used as a measure of the strength of the match. The data were used to gain insights into the kinds of cognitive processes that each test measured. For example, does the test reflect a variety of cognitive process ranging from simple process such as recall of facts and definitions or simple procedures to more complex process, or does the test focus on only a few processes such as procedures and recall of facts?

Results from the item-to-framework cross-matching of about 60 MSPAP items from six mathematics tasks appears in Table 2.6.1. The results suggest that the cognitive characteristics of the MSPAP items span the ability categories of the NAEP framework, with about a third of the MSPAP items assessing conceptual understanding, about half assessing procedural knowledge, and about one-fourth assessing problem solving. In all three cases, the degree of match between the item and the ability category was about 4 on a scale of 5 (strong) to 1 (weak), thus suggesting that across all items the match between the items and the selected categories was rather strong. These results also demonstrate that the set of MSPAP items assesses a combination of cognitive processes including understanding of concept, execution of mathematical procedures, and the ability to solve problems in new situations.

Table 2.6.1. Results from the Classification of MSPAP Items with Respect to the NAEP NAEP Categories

NAEP ABILITY CATEGORY		
Conceptual Understanding	**Procedural Knowledge**	**Problem Solving**
29% (ave. degree of match = 4.4)	48% (ave. degree of match = 4.5)	23% (ave. degree of match = 4.2)

While compiling the results from the item-to-framework cross-matching of the NAEP items according to the MSPAP, the LRDC project staff noticed that it was at times difficult to obtain a majority agreement among the panelists about the match between a NAEP item and a single MSPAP process outcome. For example, it was often the case that of eight panelists and other participants who completed this activity, four people chose the process outcome Reasoning, three chose Problem Solving, and one chose Communication. This finding was not unexpected, given that the MSPAP process outcomes are based on the cross-cutting themes from the NCTM Curriculum and Evaluation Standards (problem solving, communication, reasoning, connections), and prior research has shown that agreement among experts on which NCTM Standards theme best matches a test question is difficult to obtain (e.g., Silver & Kenney, 1994).

Because the purpose of analyzing these data was to provide preliminary information about the cognitive dimension of the MSPAP test, it was deemed sufficient to report the results in terms of a dichotomy: the percent of items judged to match at least one process outcome and the percent of items judged to match none of the four outcomes. The results appear in Table 2.6.2. Like the MSPAP items, the NAEP items spanned the MSPAP process outcomes, with about 60% matching at least one process outcome and 40% matching none of the outcomes. The degree of match between the item and the process outcome was about 3.8 on a scale of 5 (strong) to 1 (weak), thus suggesting that across all items the match between them and the selected process categories was

moderately strong. Among the four process outcomes, the two outcomes most frequently selected were problem solving and reasoning. This finding is not surprising because the items were developed according to the NAEP ability categories which include Problem Solving, a category for which reasoning is an important component. For the nearly 40% of items which the panelists and other participants could not match to a process outcome, the most frequent responses to the question concerning what the item measured were "procedures" or "understanding of concepts."

Table. 6.2.2 Results from the Classification of NAEP Items According to the MSPAP Process Outcomes

Percent of NAEP Items Matching at Least One MSPAP Process Outcome	Percent of NAEP Items Matching No MSPAP Process Outcomes
61% (ave. degree of match = 3.8)	39%

The data from this activity suggest that each test has a cognitive dimension that spans a range of processes, but due to the general nature of the ability categories from NAEP and the process outcomes from MSPAP, it was not possible to gather specific information concerning the degree to which each test measured low-level skills such as recall of facts or definitions or routine procedures or higher-level processes such as mathematical reasoning. Obtaining more information about the cognitive dimension of MSPAP and NAEP necessitated the use of other cognitive demand criteria that were more specific.

Evaluating MSPAP and NAEP Items According to Other Cognitive Demand Criteria

Design. To investigate further the kinds of cognitive demands the Maryland grade-8 test and the NAEP grade-8 test placed on students, the LRDC staff designed an activity that compared items from each test to external criteria that represent the various levels of cognitive demand. The criteria were obtained from a variety of sources such as the Evaluation standards section from Curriculum and Evaluation Standards for School Mathematics (NCTM, 1989) and other studies

involving NAEP (e.g., Romberg, Smith, Smith, & Wilson, 1992). The final set of criteria included those that represented high levels of cognitive demand (e.g., problem solving, reasoning, communication, connections, mathematical concepts, mathematical procedures) and low levels of cognitive demand (e.g., recall of facts, routine procedures, estimation). The high cognitive demand categories represent components of a student's mathematical power (NCTM, 1989).

Because of the large number of items on each test and because of the limited amount of meeting time that could be devoted to this activity during the second meeting, a subset of items from each test was selected. For MSPAP, the selected subset came from the mathematics tasks contained in one intact test booklet developed for 1996 assessment. This booklet was selected because it contained more mathematics tasks (3 tasks) and consequently more mathematics items (42 items) than were in either of the other two booklets. Although the booklet also contained tasks from other content areas such as science and social studies, the panelists worked only with the mathematics tasks. However, giving the panelists an intact booklet afforded them the opportunity to examine the test that some students actually took in 1996, thus providing additional context about the technical dimension of MSPAP as an integrated assessment taken over the course of five-days, with 1 hour and 45 minutes of testing time per day.

The subset of NAEP items came from intact blocks of items that were carefully selected to match as closely as possible certain important features of MSPAP. In particular, all MSPAP items developed for the grade-8 test were only administered to students at that grade level; all items were in constructed-response format; and all items were developed around a common theme that provided a real-world context for the mathematics concepts to be assessed. To match these three important features of MSPAP, three item blocks (containing a total of 27 items) from the main NAEP assessment and one block (containing 10 items) from a NAEP special study were selected.[3]

Looking across the three blocks from the main NAEP assessment could afford the panelists with the opportunity to examine the NAEP test as taken by the students; that is, each student in the

[3]In addition to the main NAEP assessment, NAEP periodically conducts special studies focused on areas of interest to educators. Separate blocks of items were developed for use in the special studies, and performance on these items is reported independently from performance on main NAEP.

NAEP sample worked on three blocks of NAEP items in a 45-minute time period (15 minutes per block). However, because the items in these three blocks were discrete and not connected to one another, the fourth NAEP block, called a "theme block" in 1996, was included in this activity. The theme block was similar to a MSPAP task in important ways: most but not all items were in constructed-response format, all items were administered only at grade 8, and all items were set within a common context of building a doghouse.

Looking across the four NAEP blocks, there was evidence that the characteristics of the set of items matched most of the three important characteristics of tasks within the MSPAP booklet: the majority of the NAEP items (73%) were administered only to students in grade 8, just over half of the items (54%) were in constructed-response format, and one block of items had a thematic structure that was similar to that used for the MSPAP tasks. However, choosing this particular set of NAEP items carried with it some important limitations. For example, the set was not selected to represent the distribution of items across content strands or ability categories as specified in the NAEP framework, and it did not include older items from the NAEP item pool (i.e., those items developed for the 1990 and 1992 assessments and retained to facilitate a short-term trend study of student performance). It is important to recognize these limitations here because they might affect the results from the activity.

Materials. The LRDC project staff created materials for the activity just described. Materials used by the panelists and other participants to analyze the MSPAP items appear in Appendix F.1, and those used to analyze the NAEP items appear in Appendix F.2. A copy of the Building a Doghouse theme block is in Appendix G. All of these appendices can be found in Section 3 of this report.

Procedures. The six panelists and one other participant who chose to complete this activity worked individually to evaluate each set of items according to cognitive demand categories and then summarized their impressions on the cognitive dimension of the test, as represented by the items with which they worked. They were also asked to comment on the MSPAP task as a whole and how well it provided a partial measure of a student's mathematical power. Similar comments

were elicited for the NAEP extended constructed-response questions; that is, questions on NAEP that required students to write an extended explanation of their answer.

The LRDC project staff compiled the preliminary results from this activity overnight for presentation to the panelists and other participants during the second day of the second meeting. Once they had a chance to study the results, the panelists and other participants engaged in a discussion about the results and issues concerning the cognitive dimension of the MSPAP and NAEP tests at grade 8. After the meetings were over, the LRDC staff did some additional analyses using the data from this activity.

Findings. The judgments of the panelists were analyzed according to the level of cognitive demand across all items in the MSPAP booklet and in the four blocks of NAEP items. An item was designated as having a high or low cognitive demand level based on the judgments of a simple majority of the panelists. **The results, shown in Table 2.6.3 revealed that 83% of the MSPAP items and 92% of the NAEP items were judged to have a high cognitive demand level; that is, they assessed problem solving, communication, reasoning, connections, mathematical procedures, or mathematical connections.** The remaining 17% and 8% of the MSPAP and NAEP items, respectively, were judged to have a low cognitive demand level; that is, they assessed recall of facts and definitions, routine computations, and estimation. These findings suggest that the cognitive demand of the NAEP items was slightly higher than that of the MSPAP items.

However, it is important to consider these results in light of the characteristics within the set of NAEP items selected for this activity. Recall that the four blocks of NAEP items were selected on the basis of their similarities with the MSPAP tasks with respect to grade level, item type, and thematic structure. The results in the table concerning the NAEP items reveal that without the judgments regarding the cognitive demands of the theme block items, the distribution of items according to high and low cognitive demand is more similar between MSPAP and NAEP. Thus, the items in the NAEP theme block contributed to the cognitive demand level of the set of NAEP items, and this was confirmed by the classification activity and the panelists' comments.

Results from the matching activity revealed that for the 10 items in the theme block, the majority of panelists agreed that every item matched a category representing a high level of cognitive demand.

Table 2.6.3. Summary of Results of the Analysis of Items According to Cognitive Demand

	High Cognitive Demand (% of items evaluated)		Low Cognitive Demand (% of items evaluated)	
	MSPAP	NAEP	MSPAP	NAEP
OVERALL	83	92	17	8
NAEP (without theme block)	-	89	-	11
NAEP (multiple-choice only)	-	77	-	23

Additional results in Table 2.6.3 suggested that there was evidence that item type had an effect on the distribution. In particular, considering only the NAEP multiple-choice items revealed that the MSPAP constructed-response items had a somewhat higher level of cognitive demand.

During the discussion of how well the MSPAP tasks and the NAEP extended constructed-response questions assess (at least in part) a student's mathematical power, the panelists offered these comments:

- Overall, the MSPAP tasks have the potential to provide a better measure of students' mathematical power than the extended constructed-response items in NAEP. For most of the tasks examined in this study, the set of items within a task represented the components of mathematical power (problem solving, communication, reasoning, connections, mathematical concepts and procedures) in ways that are not possible in a single NAEP extended constructed-response item to be completed in five minutes.

- However, of the four MSPAP mathematics tasks examined in this study, some were judged to contain a better mix of the components of mathematical power than other tasks, which tended to focus more on mathematical concepts and procedures and routine procedures than on problem solving and reasoning. In the case of content

areas, items that assessed topics in data analysis and statistics tended to have a particular focus on procedures (e.g., producing a circle graph; calculating a mean) and did not ask students to engage in problem solving or reasoning about the results.

- The NAEP theme block on building a doghouse was viewed favorably by the panelists and other participants. In fact, they agreed among themselves that the items within that block explored students' depth of understanding about important mathematical concepts -- especially topics of ratio and proportion -- in a very connected way that was more powerful than any of the four MSPAP mathematics tasks they examined. The doghouse theme block assessed most of the components of mathematical power, but concern was expressed that the range of content assessed was very restricted. In particular, performance on the items in this theme block would likely provide information about a student's mathematical power with respect to selected measurement and geometry concepts, but little else.

Examining Scoring Guides and Sample Student Responses for Selected MSPAP and NAEP Items

Design. The final activity involving the examination of the MSPAP and NAEP tests along the cognitive dimension was concerned with the relationship between the constructed-response items, scoring guides, and student responses that represented performance at each score level. Looking at the relationship between the items, scoring guides and responses was an important component of this study due to the fact that all MSPAP items are constructed-response as are nearly half of the NAEP items on the 1996 assessment. The underlying issue for this part of the study involved the cognitive demand level of the item and whether that demand was reflected in the scoring guide and in the sample responses. For example, if the item involves higher order thinking such as problem solving or reasoning, then the expectation is that the descriptions on the scoring guide would focus on the degree to which a response exhibited problem solving or reasoning and that the sample student responses would be reasonable examples of the score levels.

Unfortunately, the LRDC project staff realized that due to time constraints the panelists and other participants could examine only a limited number of MSPAP and NAEP items and accompanying materials. The staff selected three MSPAP items for this activity: two contiguous

items from the same task scored according to three levels and one item from a different task scored on four levels. The NAEP items were selected because their scoring schemes had the same number of levels as the MSPAP items: the NAEP regular constructed-response item was scored on three levels and the extended constructed-response item was scored on four levels.

Materials. The LRDC project staff created materials for the activity just described. An example of the packet used by the panelists and other participants to analyze the MSPAP items, scoring guides, and student responses appear in Appendix H.1, and those associated with NAEP appear in Appendix H.2. These appendices are in Section 3 of this report.

Procedures. The panelists and other participants worked individually to evaluate the items, scoring guides and sample student responses. Once they completed as much of the activity as time permitted, they engaged in a discussion about whether the cognitive demands of the MSPAP and NAEP items were sustained in the scoring guides and responses.

Findings from the examination of scoring guides and student responses. Because this activity proved to be more labor intensive than the LRDC project staff realized, most of the panelists and other participants did not have enough time to complete all parts of this activity. The consensus of the group was that either the activity should be shortened or more time should be allowed to complete it. Thus, the findings from this activity were quite limited, and only a few comments were offered.

- At least for the items examined in the activity, the cognitive demand of the MSPAP and NAEP items were both at high levels and the scoring guides appeared to reflect the demand level of the items. Some panelists posited that because the scoring guides contained more detailed descriptions for high level performance (completely correct; partially correct) and than low level performance (minimal or incorrect), the scoring scheme might differentiate better at the upper end than at the lower end of the scale.

- It was the opinion of some panelists and other participants that there were some example responses that did not match the scoring guide descriptions for performance at the particular level at which they were scored by MSPAP raters. Some responses were judged to be a better match to a lower score level; others were judged to be a

better match to a higher score level. Thus, it might be the case that for both MSPAP and NAEP, the cognitive demand of the item and scoring guide is not sustained in the scoring of the responses. This situation needs further study.

Conclusions about the Cognitive Aspects of the Tests

At the end of the discussion about the cognitive aspects of the tests, the panelists agreed on these conclusions:

- **There are only slight differences between the Maryland test and the NAEP tests at grade 8 along the cognitive dimension**. In particular and based on a subset of NAEP items that were selected to match important characteristics of the MSPAP tasks (e.g., constructed-response format; administered only to students in grade 8; common context), there was evidence that both NAEP and MSPAP measure higher level skills such as problem solving, reasoning and communication. [It is important to note that this conclusion is based on the examination of a set of NAEP items that were selected so that they would be as similar as possible to the MSPAP tasks on the basis of format, grade level at which the items were administered, and connections to a common theme. It is likely that the differences would be more pronounced if the entire NAEP item pool for grade 8 was examined.]

- **These slight differences are not sufficient to account for the magnitude of difference between proficient performance on the Maryland test and proficient performance on NAEP**.

2.7: Concluding Comments

During the latter part of the second meeting, the panelists and other participants made a final judgment about the overall congruence between the Maryland test at grade 8 and the NAEP test at grade 8. They also had the opportunity to offer any concluding comments on the tests and on the procedures used to compare them.

Overall Congruence Judgment about the Maryland Test and NAEP

Considering the technical information as presented by the LRDC project staff and their judgments about the tests along the content and cognitive dimensions, the overall consensus judgment was as follows:

> **There are differences between the Maryland test at grade 8 and the NAEP test at that grade level, but the differences are not sufficient to account for the magnitude of the difference between proficient performance on the Maryland test (48% [1994-95 results]) and on NAEP (24% [1996 results) (Musick, 1996).**

Comments on the Maryland and NAEP Tests

The focus of additional comments provided by the panelists and other participants was on the possible reasons for the discrepancy in performance at the proficient level. If the difference cannot be attributed to technical, content, or cognitive characteristics, then what are the probable contributing factors? Through group discussion they offered three possible reasons for the discrepancy: differences in stakes or consequences between MSPAP and NAEP; the connections between the test and instruction; and the standards-setting processes used.

The first reason was related to the stakes or consequences associated with the tests which might affect student motivation and performance. The results of the Maryland test are reported at school level, and individual student results are available upon request. Those schools that make substantial progress toward state standards based on the results of the test receive financial

rewards, and lack of such progress may result in reconstitution of the school. The consequences of results on the schools and availability of the individual student scores on the test may influence student motivation and performance. NAEP, on the other hand, is perceived as a low-stakes test with no consequences on schools or individual students. Hence, the consequences of the two tests may explain the differences in the reported proficiency levels.

Another reason involves the connections between the tests and instruction. The Maryland test is aligned with the curriculum, and the MSPAP learning outcomes serve as a framework for both the test and the curriculum. NAEP, however, is not tied to any curriculum. Hence the Maryland test is closely linked to the instructional experiences of the students, whereas the NAEP test is not.

A third factor was related to the standard-setting process for determining proficient performance. The proficiency level is determined on what students can do in the Maryland test. In NAEP, the proficiency level is based on aspirational criteria, that is, what students should be able to do in mathematics. Realizing that a careful examination of the definition of proficient performance and how to determine that performance was beyond the scope of the content analysis project, the participants suggested that such an examination should be the next step in the process of comparing the Maryland testing program and NAEP.

Comments on the Process Used in this Study to Compare the Tests

When given the opportunity to comment on the process as designed by the LRDC project staff and used to study the two tests, the panelists and other participants agreed that the process allowed them to examine tests in an organized fashion and in an environment that was supportive and non-threatening. They particularly liked the distributed expertise of the participants, and commented that in addition to people who knew the intricacies of the state assessment program, those who knew something about NAEP and those who can function in a "neutral" role were important to the deliberations. The participants from Maryland especially appreciated the fact that they were treated as equal partners in the process. In addition, the role of the LRDC staff members as external facilitators who created the materials and led the discussions added to the effectiveness

of the process. William Schafer, Director of the Student Assessment Division of the Maryland State Department of Education, emphasized that the role of the LRDC project staff was critical to the implementation of the process and analyses of the data. He offered the opinion that other states that decided to implement the process of looking at their assessments and NAEP along the technical, content and cognitive dimensions would do well to have people other than those from the state education department plan and implement the process.

One suggestion was related to the design of the study. Some members of the panel felt that the tests could be compared based on the difficulty of the items. Although comparison of the cognitive demands of the tests provide useful information, it does not capture the potential differences in the percent of easy or difficult items that make-up the tests. Perhaps comparing the items with respect to difficulty could be implemented in another phase of the study.

Another suggestion related to the design of the study concerned analyzing the items of the tests along the cognitive dimension. Some panelists felt that they chose more than one category for many items, and had difficulty in rank ordering their choices. Further, they commented that it was not always easy to distinguish between the categories of Reasoning and Mathematical Concepts, and between Mathematical Procedures and Computation. Since the results were reported in terms of a dichotomy (High Cognitive Demand vs. Low Cognitive Demand), it was suggested that these categories be used in the activity, with perhaps a category added for Moderate Cognitive Demand.

2.8: Reference List

BOSUN (Benchmarks of Student Understanding) Project. (1997, May). Technical guide: A working document. Pittsburgh: University of Pittsburgh.

The College Board. (1994). Mathematics framework for the 1996 National Assessment of Educational Progress. Washington, DC: National Assessment Governing Board.

Maryland State Department of Education. (1997, October). Maryland School Performance Assessment Program Learning Outcomes: Mathematics - Clarification of outcomes and suboutcomes (draft). Baltimore: Author.

Maryland State Department of Education. (1997, November). Fact sheet 20: School performance recognition awards. Baltimore: Author.

Maryland State Department of Education. (1998, May). Fact sheet 5: School reconstitution: State intervention procedures for schools not progressing toward state standards. Baltimore: Author.

Maryland State Department of Education, CTB/McGraw Hill, & Measurement Incorporated (1996, December). Technical report - Pre release edition - 1996 Maryland School Performance Assessment (MSPAP). Baltimore: Authors.

Mullis, I. V. S., Dossey, J. A., Owen, E. H., & Phillips, G. W. (1991, June). The STATE of mathematics achievement: NAEP's 1990 assessment of the nation and the trial assessment of the states. Washington, D. C.: National Center for Education Statistics.

Musick, M. D. (1996, June). Setting educational standards high enough. Atlanta: Southern Regional Education Board.

National Council of Teachers of Mathematics. (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: Author.

Romberg, T. A., Smith, M., Smith, S., & Wilson, L. (1992, June). The feasibility of using international data to set achievement levels for the National Assessment of Educational Progress (NAEP). Madison, WI: National Center for Research in Mathematics Education.

Silver, E. A., & Kenney, P. A. (1994). The content and curricular validity of the 1992 NAEP Trial State Assessment (TSA) in mathematics. In National Academy of Education, The Trial State Assessment: Prospects and realities - Background studies (pp. 231-284). Stanford, CA: The Academy.

Zawojewski, J. S., & Silver, E. A. (1998). Assessing conceptual understanding. In G. W. Bright & J. M. Joyner (Eds.), Classroom assessment in mathematics: Views from a National Science Foundation Working Conference (pp. 287-295). Lanham, MD: University Press of America.